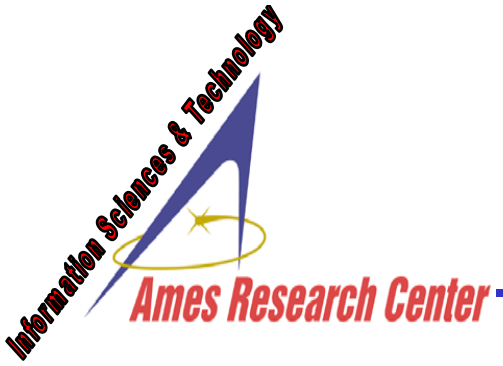
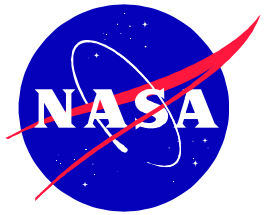
Information Sciences & Technology

Ames Research Center

# Finding a Needle in the Haystack: Discovering Recurring Anomalies *Described* in Text Documents

Ashok N. Srivastava, Ph.D.
Intelligent Data Understanding Group Leader
NASA Ames Research Center



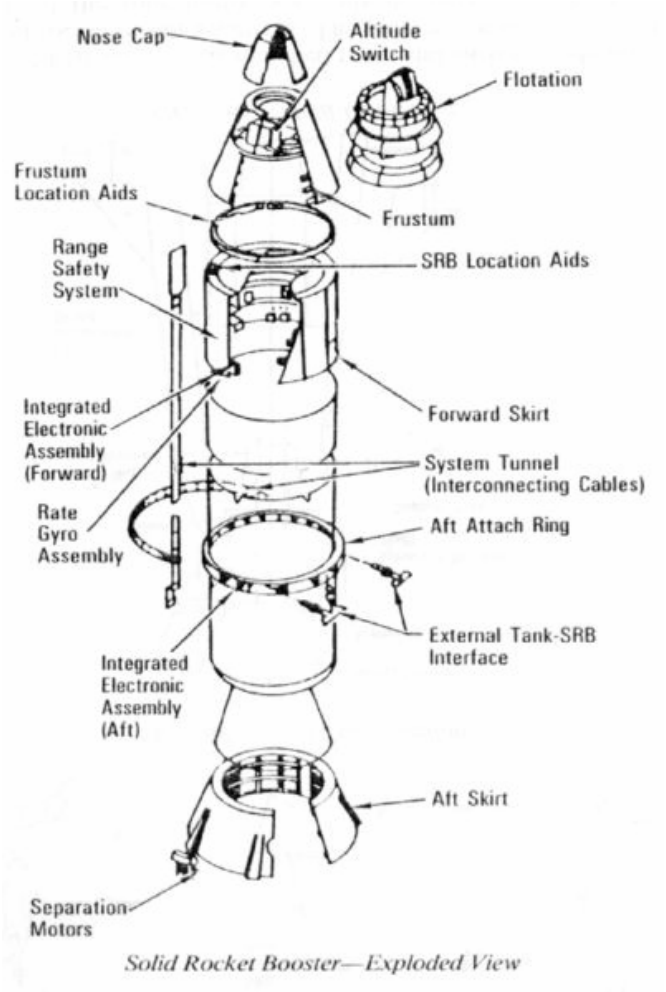

**Recurring Anomaly Detection System (ReADS) Team**

Ashok Srivastava, Ph.D.
Dawn McIntosh
Pat Castle
Manos Pontikakis
Vesselin Diev
Brett Zane-Ulman
Eugene Turov
Ram Akella, Ph.D.
Zuobing Xu
Sakthi Preethi Kumaresan
Tom Ferryman, Ph.D.
Advance Engineering Network Team

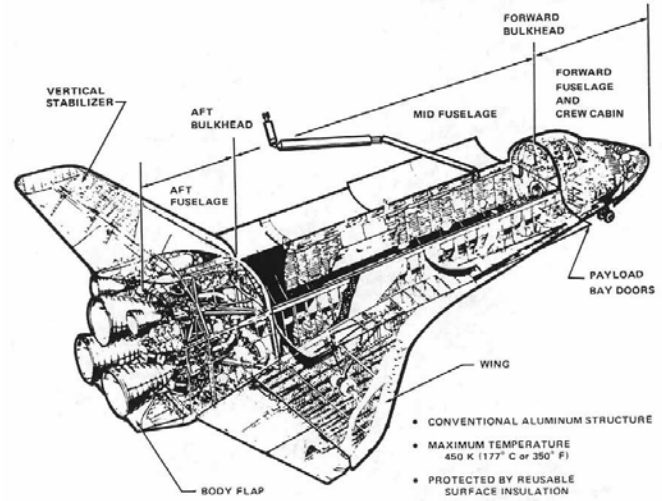
*Team Members are NASA, Contractors and Students*

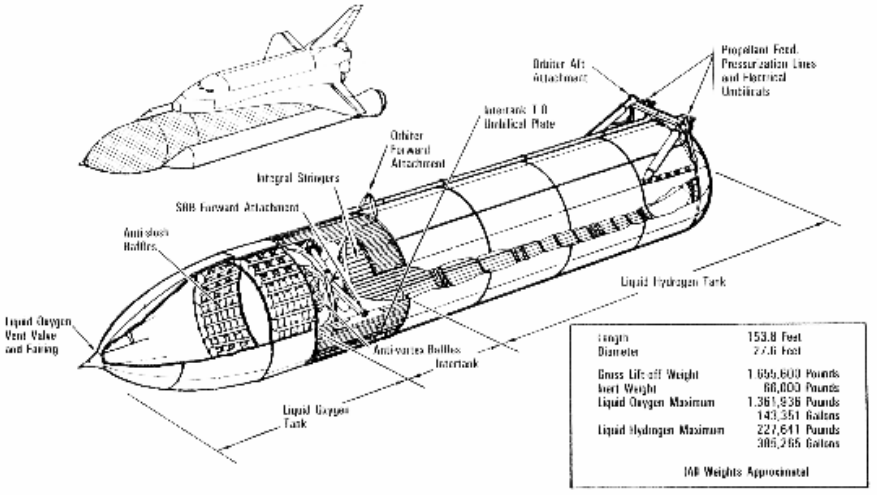
*All schematic diagrams and pictures in this presentation are publicly available on the internet.*

# Some Systems we are Considering




*Solid Rocket Booster—Exploded View*



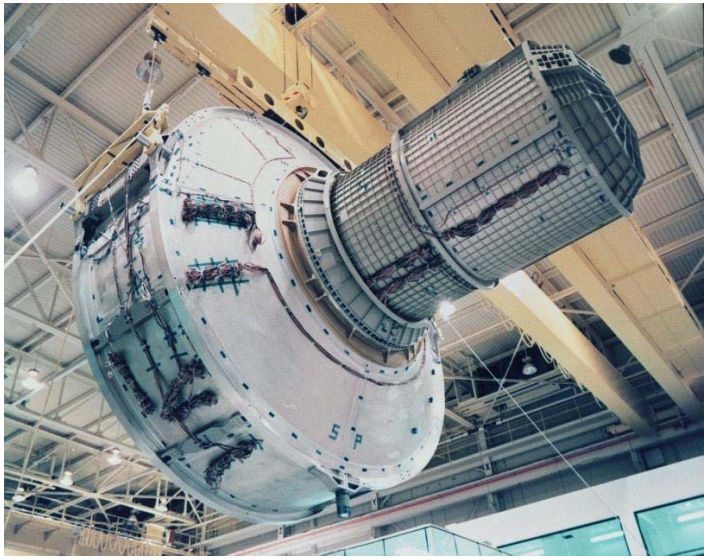



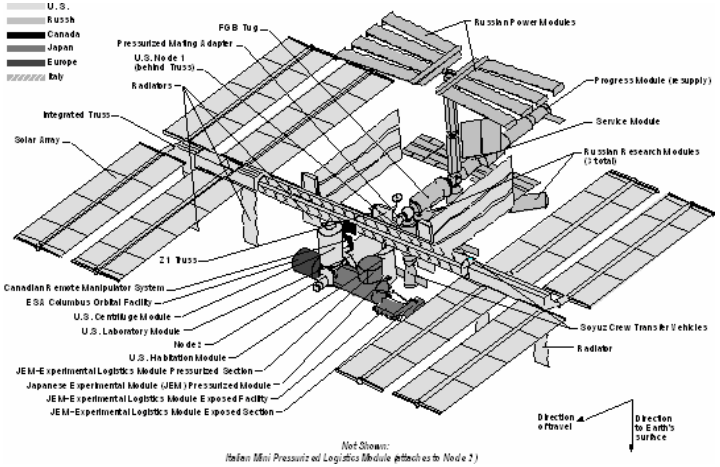

*Lightweight External Tank*

Information Sciences & Technology

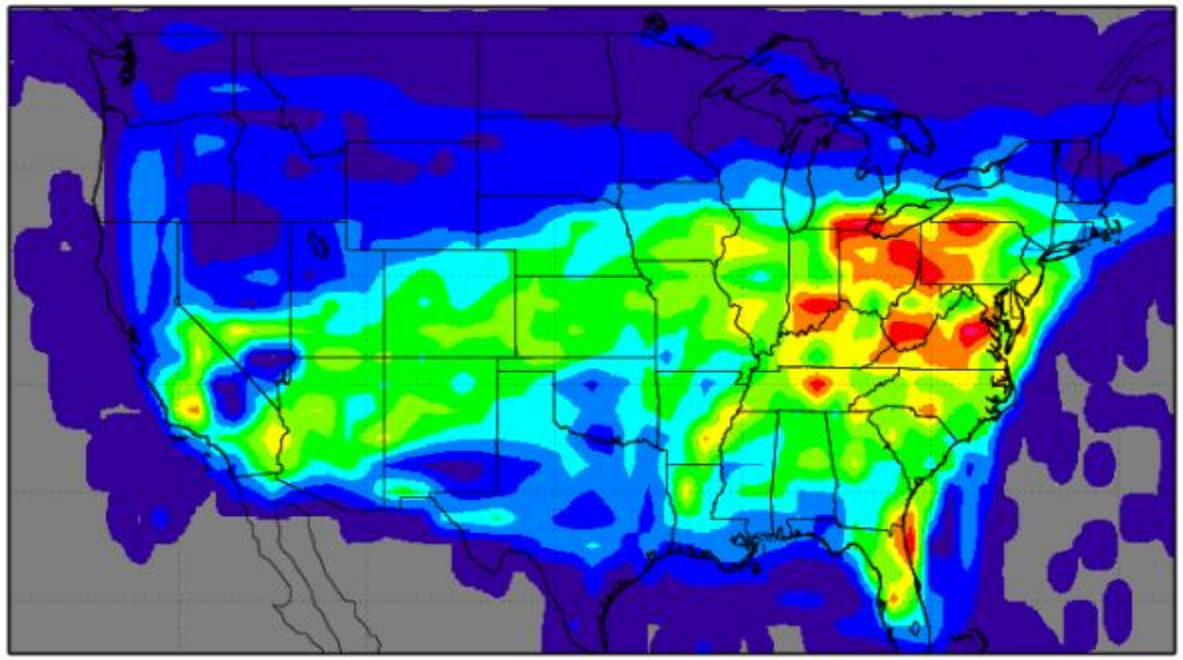
Ames Research Center

# International Space Station





# The National Air Space



# Problem Reports

- Each system can have several hundred-thousand reports written about them.
- Hundreds or thousands of different authors.
- In some cases, different languages are used.
- Reports can be 0.5-4 pages long.
- Each system has its own set of acronyms
- These systems have been around for decades and are continuously being modified.
- Each author has his or her own perspective.

**NASA Acronyms**

[A][B][C][D][E][F][G][H][I][J][K][L][M][N][O][P][Q][R][S][T][U][V][W][X][Y][Z]

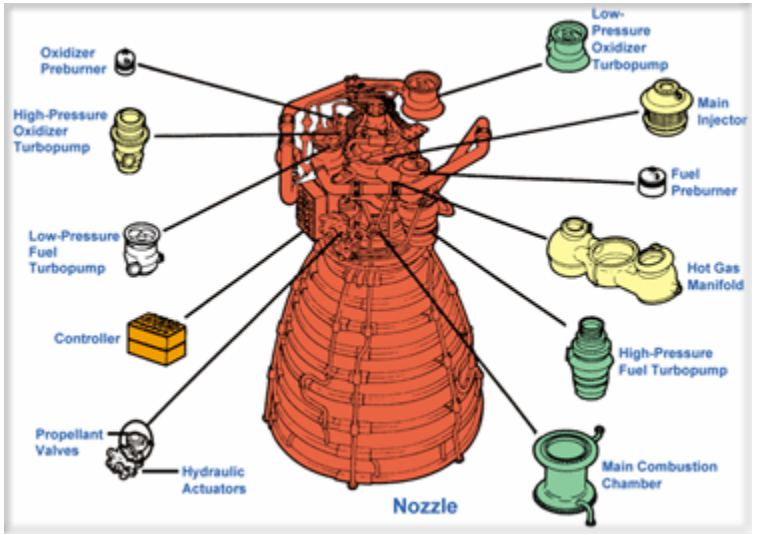
Did you ever wonder what those strange conglomerations of letters meant? Here's your chance to find out what those NASA acronyms stand for. About Space and Astronomy presents the Guide to NASA Acronyms. From Navigation and Guidance to Network Interface Data System - NASA Management Instruction to Nose to Z-Axis, you'll find them here. This page lists acronyms starting with N. For others, click on the appropriate letter above or below.


[N]

N: Neutrons
N: Newton
N: North
N&G: Navigation and Guidance (G&N preferred)
N/O: Normally Open
N/A: Not Applicable
N/A: Next Assembly
N/B: Narrow Band
N/C: Numerical Control
N/C: Normally Closed
N/C: Nose Cone
N/C: Not Critical
N/D: Need Date
N/P: Not Provided
N/W: Network
N02: Nitrogen Dioxide
N2: Nitrogen
N2: Nitrogen
N204: Nitrogen Tetroxide
N2H4: Hydrazine
N2HO4: Nitrogen Peroxide
N2O4: Nitrogen tetroxide
NA: Next Action
NA: Not Applicable
NAAL: North American Aerodynamic Laboratory (Wind Tunnel)
NAC: Nacelle
NAEC: Naval Air Engineering Center
NAM: National Association Of Manufacturers
NAP: Navigation Analysis Program
NAR: Numerical Analysis Research
NARS: National Archives and Record Services
NAS: National Aircraft Standard
NAS: National Academy of Sciences
NAS: Naval Air Station
NASA: National Aeronautics and Space Administration
NASA: National Aeronautics and Space Administration
NASCOM: NASA Communications (Network)
NASTRAN: NASA Structural Analysis
NATF: Naval Air Test Facility
NATL: National
NAV: "Navigate, Navigation"
NAVAID: Navigation Aid
NAVDAD: Navigationally Derived Air Data
NAVPOOL: Navigation Parameter Common Pool
NAVSAT: Navigation Satellite
NB: Navigation Base
NB: No-Bias (Relay)
NB: Nitrogen Base
NB: Narrow Band
NBF: Neutral Buoyancy Facility
NBS: National Bureau of Standards
NBT: Neutral Buoyancy Trainer

# Problem Definition

- Develop a system that will automatically discover recurring anomalies given a stack of 100,000+ reports.
- Some types of reports are pre-classified into anomaly categories. Others are not classified into categories.
- A recurring anomaly is a reported problem that happens more than once regarding:
  - The same system
  - Similar systems
  - Functionally related systems



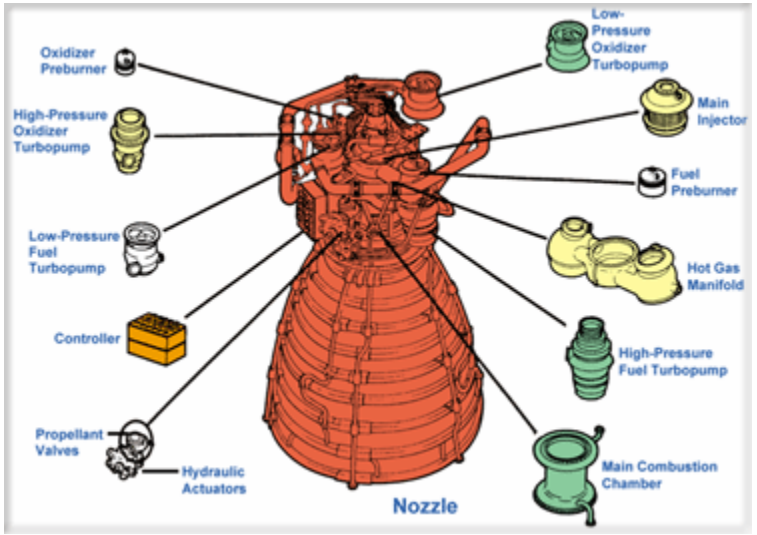
# Fingerprints of Anomalies

**Identifiable Anomalies**

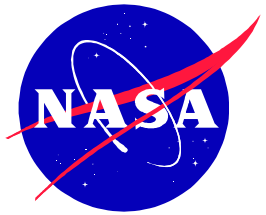
- Recurrent failures
- Problems that cross traditional system boundaries so failure effects are not fully recognized
- Problems that have been accepted by repeated waivers
- Discrepant conditions repeatedly accepted by routine analysis
- Events with unknown causes.

**Hard to Identify**

- Single failures
- Identification of the root cause of anomalies that propagate through several systems.

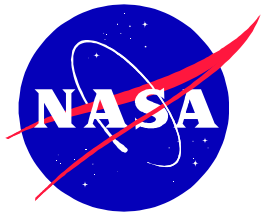


# Why not just…




- Have people read the reports and come up with recurring anomalies?
- Use Google?
- Do keyword search?
- Generate good forms to collect data to enable discovery of anomalies?
- Focus on sensor data?

# Types of Reports




- **Problem Reporting And Corrective Action (PRACA)**
  - Usually for engineering systems such as Shuttle, ISS
  - Usually have sections describing each element
  - *Usually* are not pre-assigned into anomaly categories.
  - Written by engineers and scientists
- **Aviation Safety Reporting System (ASRS)**
  - Publicly available safety reports regarding commercial airliners.
  - Are categorized into one of 62 anomaly categories.
  - Written by pilots, crew, maintenance
- **Aviation Safety Action Program (ASAP)**
  - Proprietary safety reports for major airlines (AA, UAL, etc.)
  - Are not pre-categorized
  - Written by pilots, crew, maintenance

# Summary of Approach

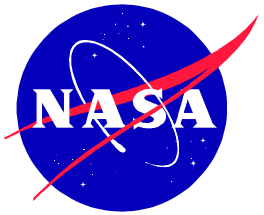
**Approach for PRACA data**

- These reports *do not* have an anomaly category
- We first perform von-Mises Fisher clustering to break the corpus into groups.
- Develop similarity measures.
- Use an agglomerative clustering technique to link documents.
- Documents linked early may be recurring anomalies.
- Link documents that reference each other.

**Approach for ASRS data**

- These reports *are* preclassified into anomaly categories.
- Develop Natural Language Processing (NLP) techniques to identify terms and phrases related to anomalies.
- Build classifier(s) to learn mappings from documents to categories.
- Test quality of classifiers.
- Apply to new corpora from airlines that have no categorization.

# von-Mises Fisher Clustering



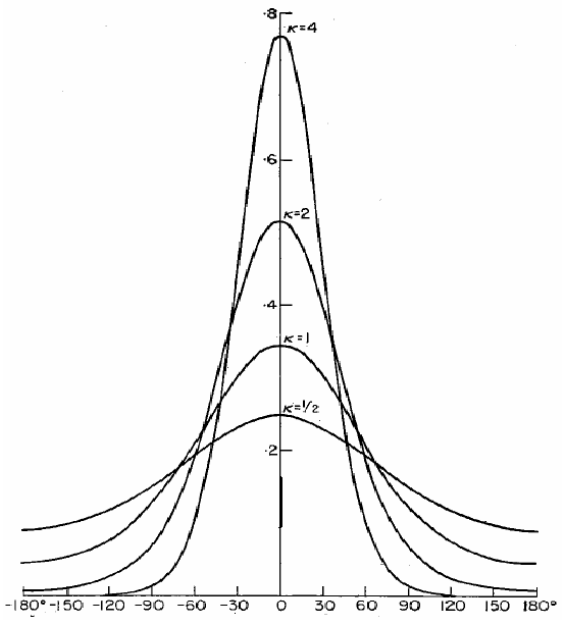

- Innovation by Banerjee, Dhillon, Ghosh, and Sra 2005.
- Idea: convert document vectors into directional data by normalizing to unit length.
- Create a generative model for a d-dimensional document vector x of unit length:

$$f(\mathbf{x}|\boldsymbol{\mu},\kappa) = c_d(\kappa)e^{\kappa\boldsymbol{\mu}^T\mathbf{x}},$$

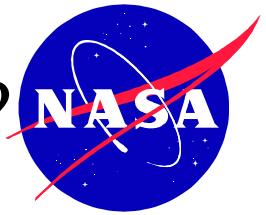
$$c_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}I_{d/2-1}(\kappa)}$$

Bessel function of the first kind, order d/2-1

Density of the von Mises distribution for $\mu_0 = 0^0$ and $\kappa = \frac{1}{2}, 1, 2, 4$.



# Why is VMF a good model for text?




- Document vectors are $L_2$ normalized to make them unit norm.
- Assumption: Direction of documents is sufficient to get good clusters.
- Two documents - one small, one lengthy - on the same topic will have the same direction and hence put in the same cluster.
- This unit normalized data lives on a sphere in a $R^{(d-1)}$ dimensional space.

# Connections with the Normal Distribution

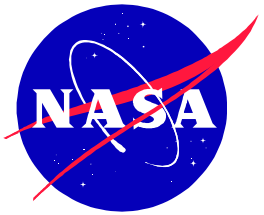
- A circular random variable ‘θ’ follows a von Mises Distribution if its pdf is given by:

$$g(\theta;\mu_o,\kappa) = \frac{I}{2\pi I_o(\kappa)} \exp \kappa \cos(\theta-\mu_o),$$
$$0 \le \theta \le 2\pi, \kappa > 0, 0 \le \mu_o \le 2\pi$$

For large K the random variable ‘θ’ is distributed as $N(\mu o, 1/K^{1/2})$

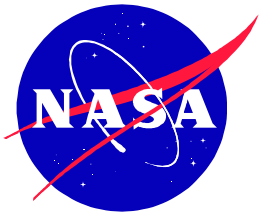
Relation to Bivariate Normal Distribution:
Let x and y be independent normal variables with means (cos $\mu o$ , *sin* $\mu o$ ) and equal variances 1/ K. The p.d.f. of the polar variables (r, θ) is VMF The conditional distribution of θ for r = 1, is the VMF($\mu o$, K).

# Connections with Normal II




- Maximum Entropy Characterization:
  - Given a fixed mean and variance the Gaussian is the distribution that maximizes the entropy.
  - Likewise given a fixed circular variance $\rho$ and mean direction $\mu_o$, the VMF distribution maximizes the entropy.
- Central Limit Theorems
  - For data on a line, the CLT says that the Normal is the limiting distribution.
  - Whereas for directional data, the limiting distribution of the sum of 'n' independent random variables is given by the Uniform Distribution.

# Connections with the Normal Distribution



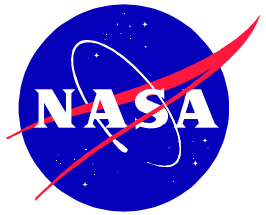

Unfortunately there is no distribution for directional data which has all properties analogous to the linear normal distribution. The VMF has some but not all of these desirable properties.

The VMF provides:

- simpler ML estimates.
- tractable distribution in hypothesis testing.

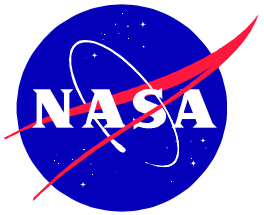
See Banarjee et. al. for details of the maximum likelihood estimates and the EM derivation.

# Similarity measures



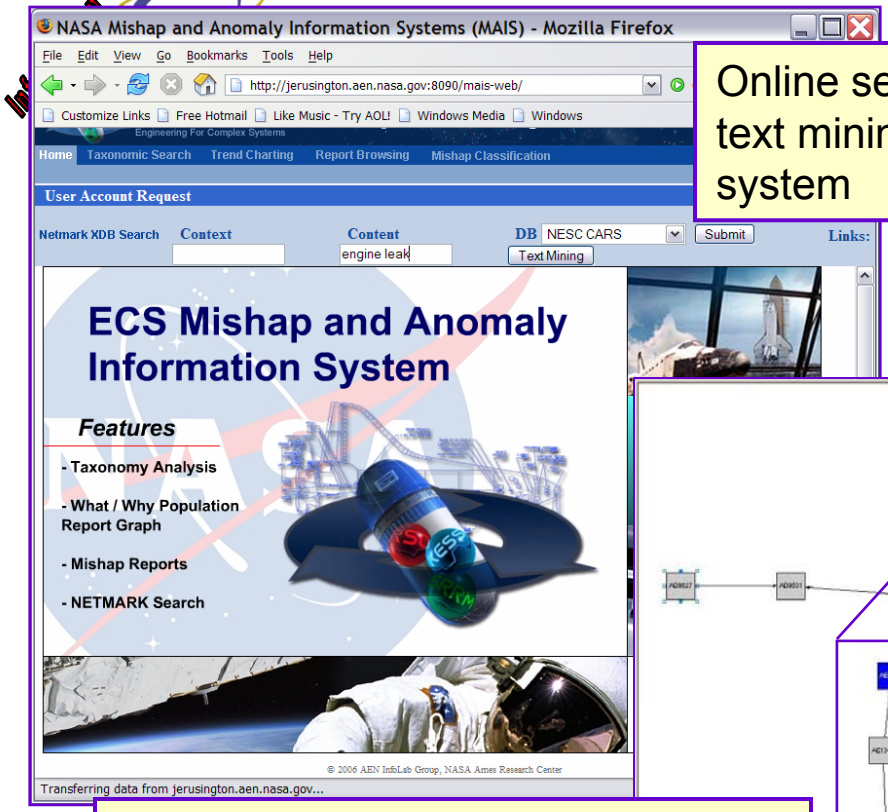

- The vMF distribution implies that the cosine measure betweeen documents is a natural similarity measure.
- We tested numerous other measures by assuming a language model (see Srivastava et. al., 2005) and measuring the distance between the distributions of words using Kullback-Leibler.
- All methods performed identically within the error bars.

# Discovering Recurring Anomalies

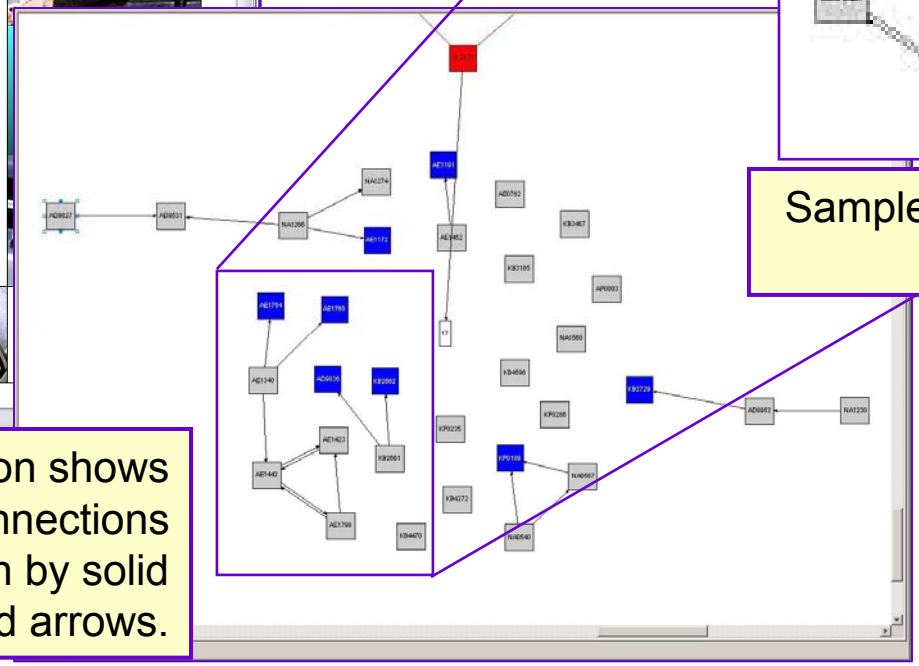


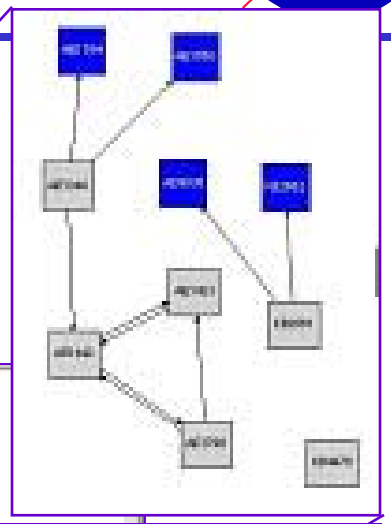

- After calculating the distance between each document, the algorithm applies single linkage, i.e., nearest neighbor, to create a hierarchical tree representing connections between documents.
  - Also generates an 'inconsistency coefficient' which is a measure of the relative consistency of each link in the tree.
- The hierarchical tree is partitioned into clusters by setting a threshold on the inconsistency coefficient.
  - A high inconsistency coefficient implies that the reports could be very different and still be sorted into the same cluster.
- Currently the inconsistency coefficient threshold is set very low, which returns many smaller clusters of very similar reports.
  - Clusters of single documents are excluded from the recurring anomaly results.

# ReADS System & Interactive Visualization




Online search & text mining system





Sample Recurring Anomalies

ReADS visualization shows documents as boxes. Connections between reports are shown by solid lines and arrows.

Recurring Anomaly Detection System- ReADS

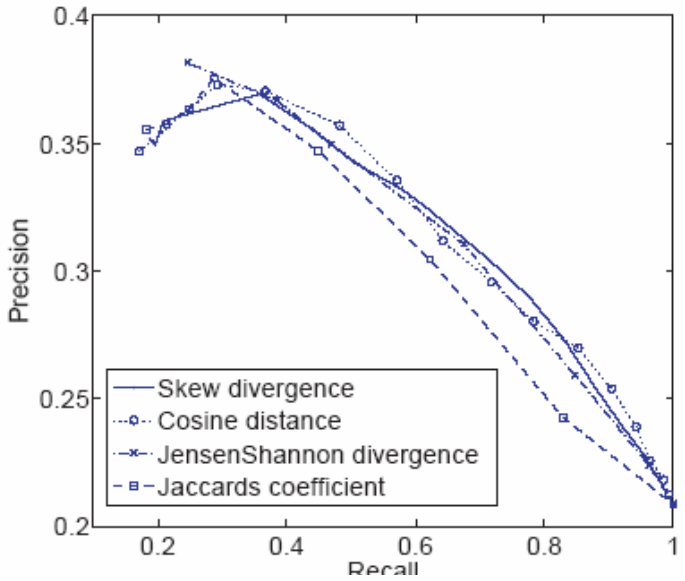
# How well does the system work?



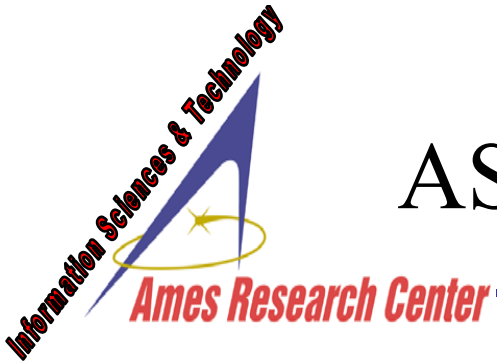

- We compared the results of our system against a set of recurring anomalies identified by humans on a sample data set of nearly 7400 reports.
- We discovered many recurring anomalies that were missed by the experts.
- We missed anomalies and also had a relatively large false positive rate.

$$Precision = \frac{R^+}{R^+ + N^+}$$

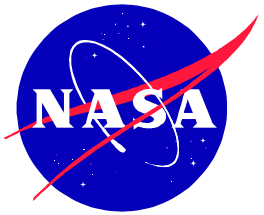
$$Recall = \frac{R^+}{R^+ + R^-}$$




| | Labeled by Expert as recurring anomaly | Not labeled by Expert as recurring anomaly |
|---|---|---|
| Detected as recurring anomaly | $R^+$ | $N^+$ |
| Not detected as recurring anomaly | $R^-$ | $N^-$ |




# ASRS Anomaly Classification




- These reports are already coded into 60 overlapping categories.
- We developed Natural Language Processing techniques to preprocess this data before submission to SVM for classification.

# Sample Language Normalization & Term Reduction

JUST PRIOR TO TOUCHDOWN, LAX TWR TOLD US TO GO AROUND BECAUSE OF THE ACFT IN FRONT OF US. BOTH THE COPLT AND I, HOWEVER, UNDERSTOOD TWR TO SAY, 'CLRED TO LAND, ACFT ON THE RWY.' SINCE THE ACFT IN FRONT OF US WAS CLR OF THE RWY AND WE BOTH MISUNDERSTOOD TWR'S RADIO CALL AND CONSIDERED IT AN ADVISORY, WE LANDED. AS WE TAXIED TO THE GATE, TWR REQUESTED THAT I CALL THEM FROM A PHONE WHEN I HAD THE OPPORTUNITY (I CALLED FROM THE GATE). IT WAS ON THE PHONE THAT I DISCOVERED TWR HAD SENT US AROUND. IN HINDSIGHT, FROM THEIR PERSPECTIVE, GOING AROUND WAS THE PRUDENT THING TO DO. I HAVE BECOME TOO CONDITIONED IN THE PAST FEW YRS IN BEING VECTORED INTO A VISUAL APCH BEHIND AN ACFT THAT IS TOO CLOSE. REGRETTABLY, IN THIS SIT, CONFUSION AND MISUNDERSTANDING PUT US IN A DIFFICULT SIT.

↓ Expand Acronyms, Simplify Punctuation ↓

JUST PRIOR TO TOUCHDOWN, LAX tower TOLD US TO GO AROUND BECAUSE OF THE aircraft IN FRONT OF US. BOTH THE copilot AND I, HOWEVER, UNDERSTOOD tower TO SAY, clear TO LAND, aircraft ON THE runway. SINCE THE aircraft IN FRONT OF US WAS clear OF THE runway AND WE BOTH misunderstand tower RADIO CALL AND CONSIDERED IT AN ADVISORY, WE LANDED. AS WE TAXIED TO THE GATE, tower REQUESTED THAT I CALL THEM FROM A PHONE WHEN I HAD THE OPPORTUNITY I CALLED FROM THE GATE. IT WAS ON THE PHONE THAT I DISCOVERED tower HAD SENT US AROUND. IN HINDSIGHT, FROM THEIR PERSPECTIVE, GOING AROUND WAS THE PRUDENT THING TO DO. I HAVE BECOME TOO CONDITIONED IN THE PAST FEW year IN BEING VECTORED INTO A VISUAL approach BEHIND AN aircraft THAT IS TOO CLOSE. REGRETTABLY, IN THIS situation, CONFUSION AND MISUNDERSTANDING PUT US IN A DIFFICULT situation.

↓ Stemming, Remove Non-Informative Terms, Phrasing ↓

PRIOR _ TOUCHDOWN _ tower TOLD _ _ goaround _ _ _ aircraft _ FRONT _ _ _ _ copilot _ _ _ understand tower _ SAY clear _ LAND aircraft _ _ runway _ _ aircraft _ FRONT _ _ _ clear _ _ runway _ _ _ misunderstand tower RADIO CALL _ consider _ _ advise _ lan _ _ taxiedto _ GATE tower request _ _ CALL _ _ _ PHONE _ _ _ _ OPPORTUNITY _ call _ _ GATE _ _ _ _ PHONE _ _ discover tower _ SENT _ _ _ HINDSIGHT _ _ PERSPECTIVE go _ _ _ prudentthing _ _ _ _ _ _ condition _ _ PAST _ year _ _ vector _ _ VISUAL approach _ _ aircraft _ _ _ CLOSE REGRETTABLY _ _ situate confuse _ misunderstand PUT _ _ _ difficultsituation

# Report before language normalization

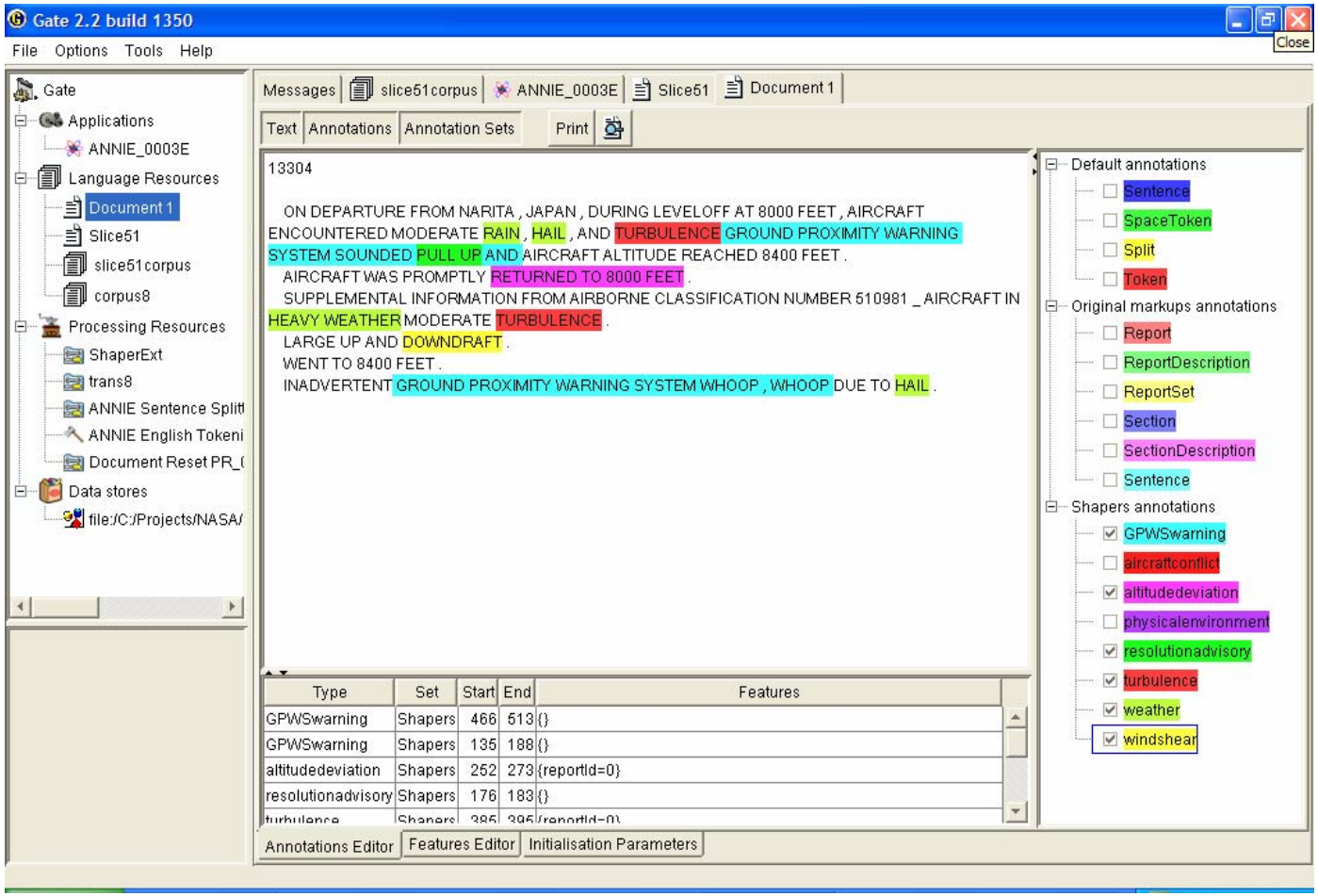
- ON DEP FROM NARITA, JAPAN, DURING LEVELOFF AT 8000 FT, ACFT ENCOUNTERED MODERATE RAIN, HAIL, AND TURB (GPWS SOUNDED 'PULL UP) AND ACFT ALT REACHED 8400 FT. ACFT WAS PROMPTLY RETURNED TO 8000 FT. SUPPLEMENTAL INFO FROM ACN 510981: ACFT IN HVY WX/MODERATE TURB. LARGE UP- AND DOWN-DRAFTS. WENT TO 8400 FT. INADVERTENT GPWS 'WHOOP, WHOOP' DUE TO HAIL.

# After language normalization

Weather
Severe Weather
Turbulence
Ground Proximity Warning System activation
Resolution Advisory
Altitude Deviation
Windshear

ON DEPARTURE FROM NARITA , JAPAN , DURING LEVELOFF AT 8000 FEET , AIRCRAFT ENCOUNTERED MODERATE RAIN , HAIL , AND TURBULENCE GROUND PROXIMITY WARNING SYSTEM SOUNDED PULL UP AND AIRCRAFT ALTITUDE REACHED 8400 FEET . AIRCRAFT WAS PROMPTLY RETURNED TO 8000 FEET . SUPPLEMENTAL INFORMATION FROM AIRBORNE CLASSIFICATION NUMBER 510981 _ AIRCRAFT IN HEAVY WEATHER MODERATE TURBULENCE . LARGE UP AND DOWNDRAFT . WENT TO 8400 FEET . INADVERTENT GROUND PROXIMITY WARNING SYSTEM WHOOP , WHOOP DUE TO HAIL .|

# Natural Language Processing



# Natural Language Processing (II)

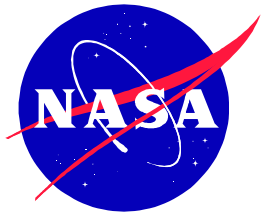
- ATC Communication Anomaly
- Altitude Deviations
- Airspace Violations
- Approach Anomalies
- Controlled Flight Towards Terrain
- Equipment Problem
- Fire
- Fuel
- GPWS
- Ground Encounter
- Ground Excursion
- Ground Incursion
- Hazardous Materials Violation
- In-flight Encounters
- Landing Anomalies
- Loss of Control
- Maintenance Problem
- Near Miss
- Passenger/Cabin Event
- Speed Deviation
- Takeoff Anomalies
- TCAS
- Turbulence
- Unstabilized Approach
- Weather
- Windshear

**Over 200 building block concepts mapped to 39 Major Categories**

# Examples of NLP predictions for "Turbulence" Anomaly

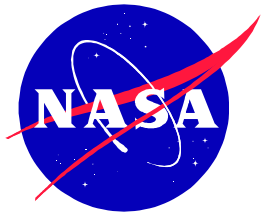
- THE AIR WAS VERY TURBULENT , THE FREEZING LEVEL WAS AT 10000 FEET MEAN SEA LEVEL , WIDESPREAD SHOWERS OBSCURED THE MOUNTAINS NORTH OF THE TEMPORARY FLIGHT RESTRICT , AND WE HAD INSUFFICIENT FUEL TO RETURN TO MCCALL , IDENTIFY MLY , TO LAND .
- WE HAD JUST FINISHED WITH OUR FIRST BEVERAGE SERVICE , WHEN HEAVY TURBULENCE HIT .
- BECAUSE OF THE TURBULENCE , MY HAND INADVERTENTLY HIT THE VOLUME CONTROL AND LOWERED THE VOLUME TO AN INAUDIBLE LEVEL WITHOUT DETECTION BY THE CREW .
- COULD NOT MAINTAIN VISUAL METEOROLOGICAL CONDITIONS IN THE TOPS AND STARTED TO GET LIGHT PRECIPITATION AND LIGHT TO MODERATE CHOP .
- A PASSENGER VIEWING THE MANDATORY VIDEO BEFORE FLIGHT FULLY UNDERSTANDS THE RESULTS OF NOT WEARING A SEATBELT IF SHOWN POSSIBLE SCENARIOS INCLUDING UNEXPECTED TURBULENCE AND STOPPING SHORT ON THE RUNWAY DURING TAXI .


# Example of False Negative




I PARKED BEHIND THE HOLD SHORT LINE AT THE RUNUP AREA OF RUNWAY 18 AND PROCEEDED TO COMPLETE THE BEFORE TAKEOFF CHECKLIST . AFTER COMPLETING THE BEFORE TAKEOFF CHECKLIST , I TAXIED THE AIRPLANE INTO A POSITION TO CLEAR THE BASE AND THE FINAL APPROACH PATH FOR INCOMING TRAFFIC . NO TRAFFIC WAS OBSERVED ON THE FINAL OR BASE . ADDITIONALLY , NO RADIO CALL WAS MADE BY THE INCOMING AIRCRAFT . NOT SEEING OR HEARING ANY AIRCRAFT ON A FINAL APPROACH , I PROCEEDED TO CROSS THE HOLD SHORT LINE AND TAXIED ONTO THE RUNWAY . AS I ALIGNED THE AIRCRAFT WITH THE RUNWAY CENTERLINE , AN AIRCRAFT FLEW OVER MINE AND EXECUTED A MISSED APPROACH .

# Raw Text & Language Normalization



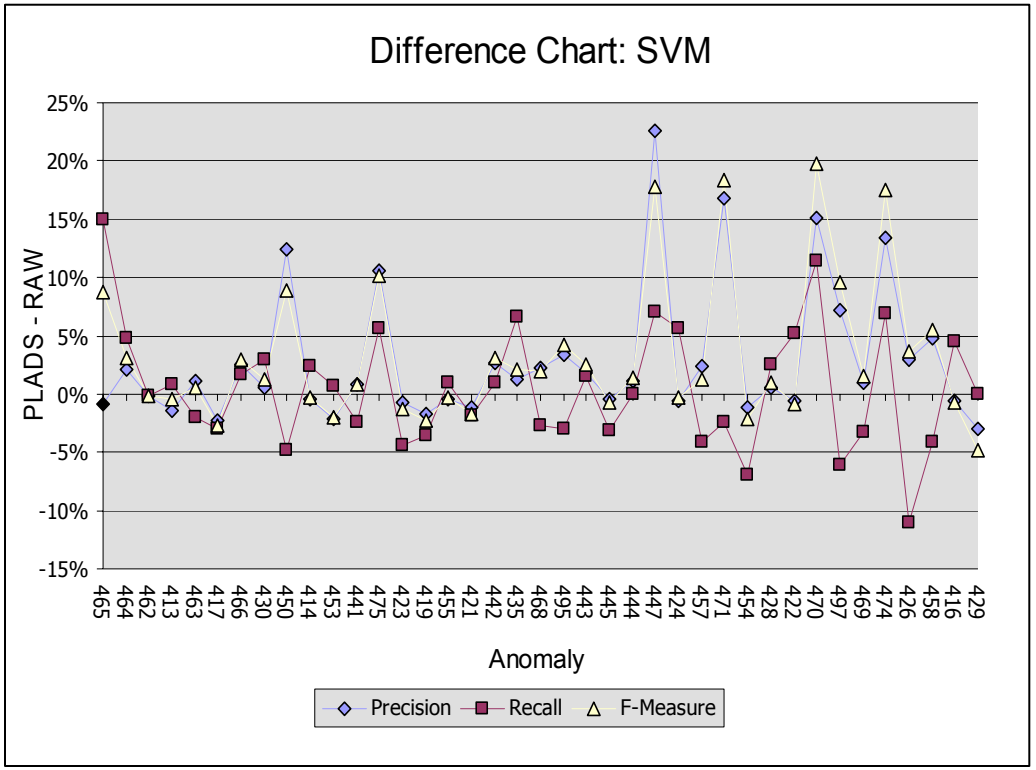

In order to classify the documents, they are first formatted into a document-term frequency matrix. The cells of the matrix are the frequency count of the terms that appear in the document.

| | Term 1 | Term 2 | Term 3 | Term 4 |
|---|---|---|---|---|
| Document 1 | 0 | 1 | 0 | 4 |
| Document 2 | 0 | 3 | 0 | 0 |
| Document 3 | 2 | 8 | 1 | 0 |

- PLADS reduced the total number of terms in 27000 documents from 44940 to 31701
- PLADS reduced classification computation time by 0%-10%

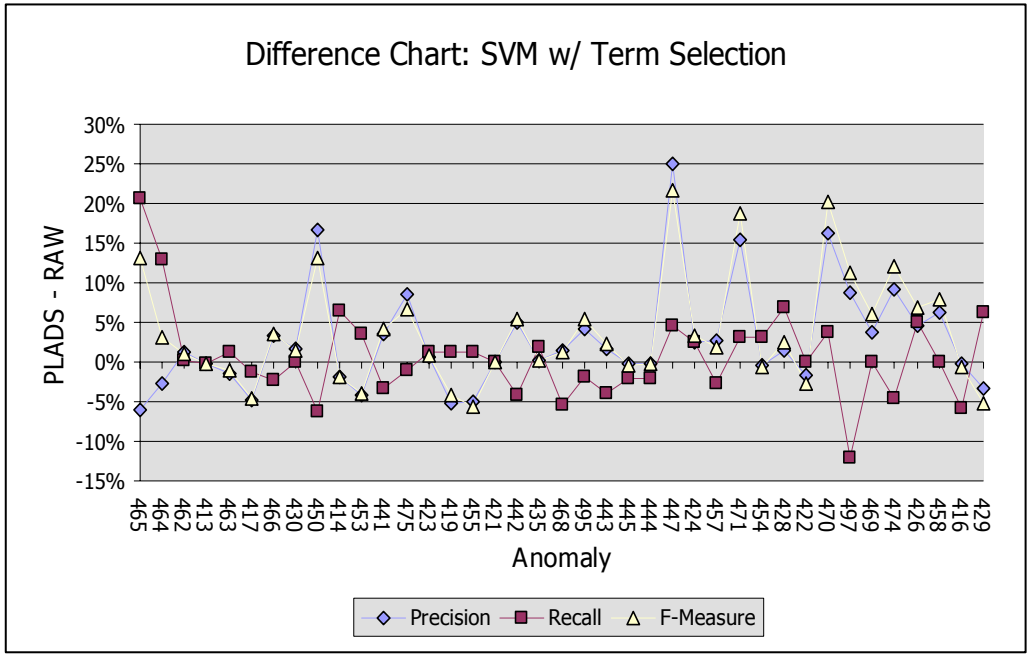
# Comparison of Raw Text vs. Language Normalization using SVM

- All terms used, no additional term reduction applied
- Language Normalization improves precision 2% on average
- Language Normalization improves recall 2% on average



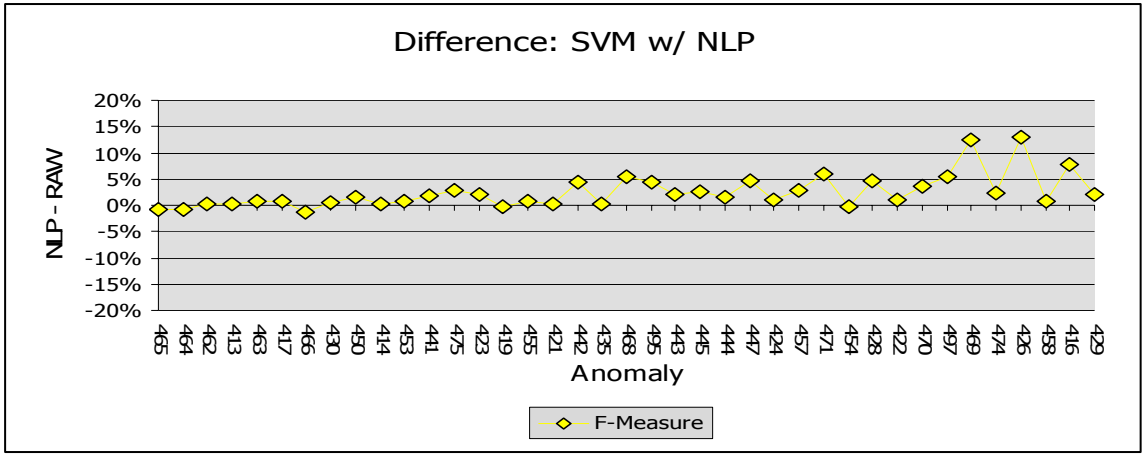
# Comparison of Raw Text vs. LN with Terms Selection

- 1000 terms selected using Information Gain
- LN improves precision 2% on average
- LN improves recall 3% on average

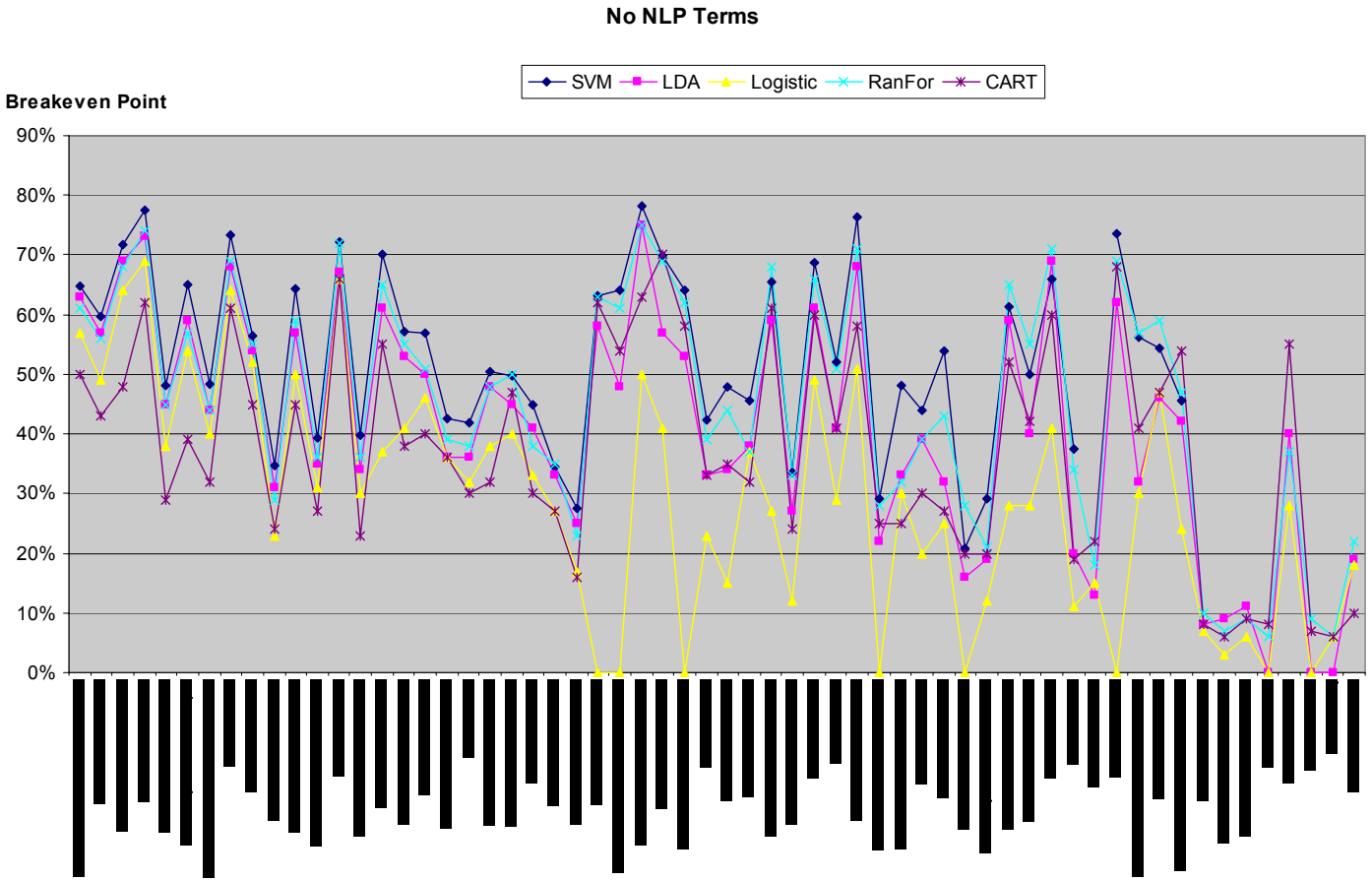


# Comparison of Raw Text vs. NLP with Terms Selection

- 500 terms selected using Information Gain
- NLP improves F-measure 3% on average

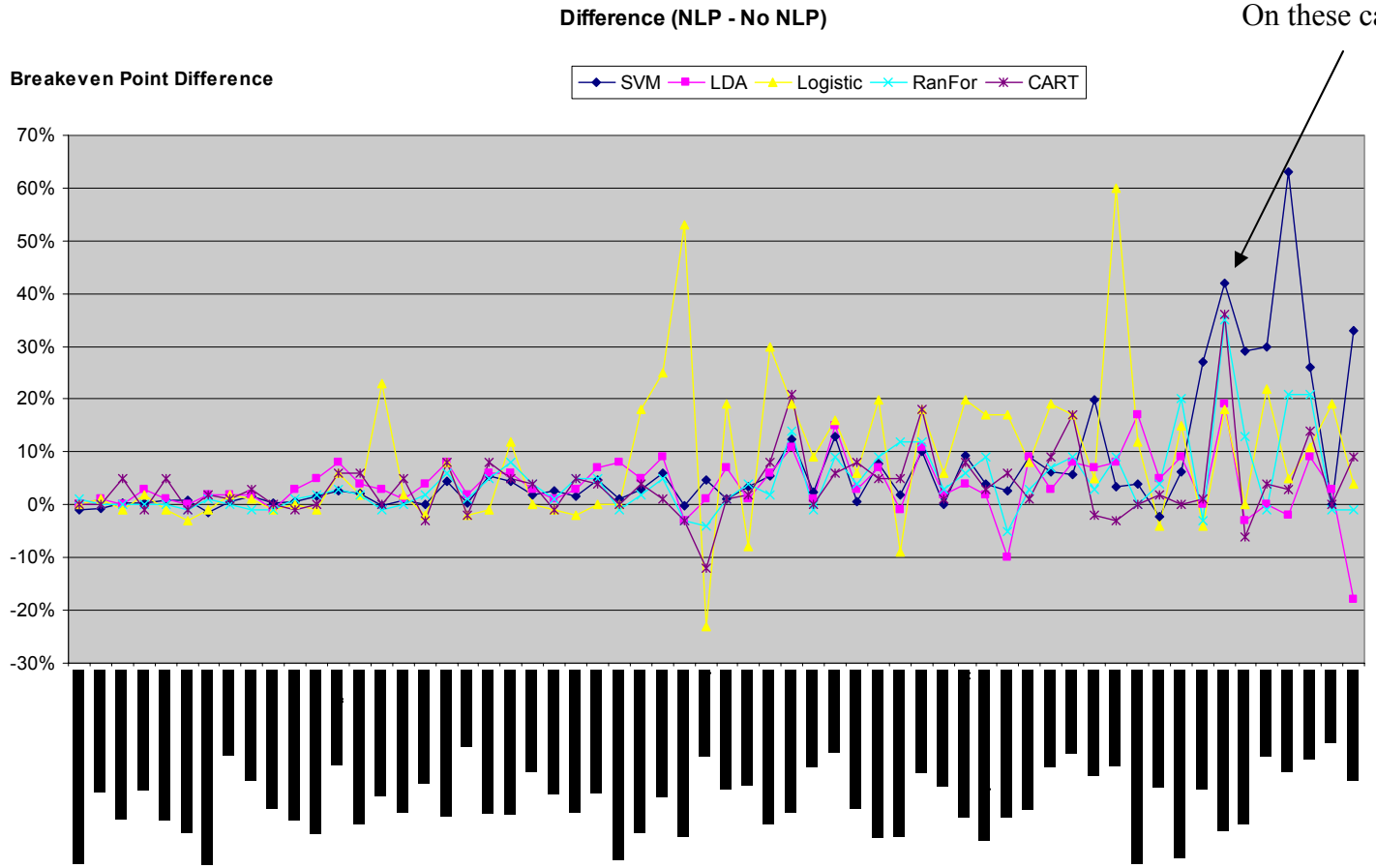


# Overall Results without NLP

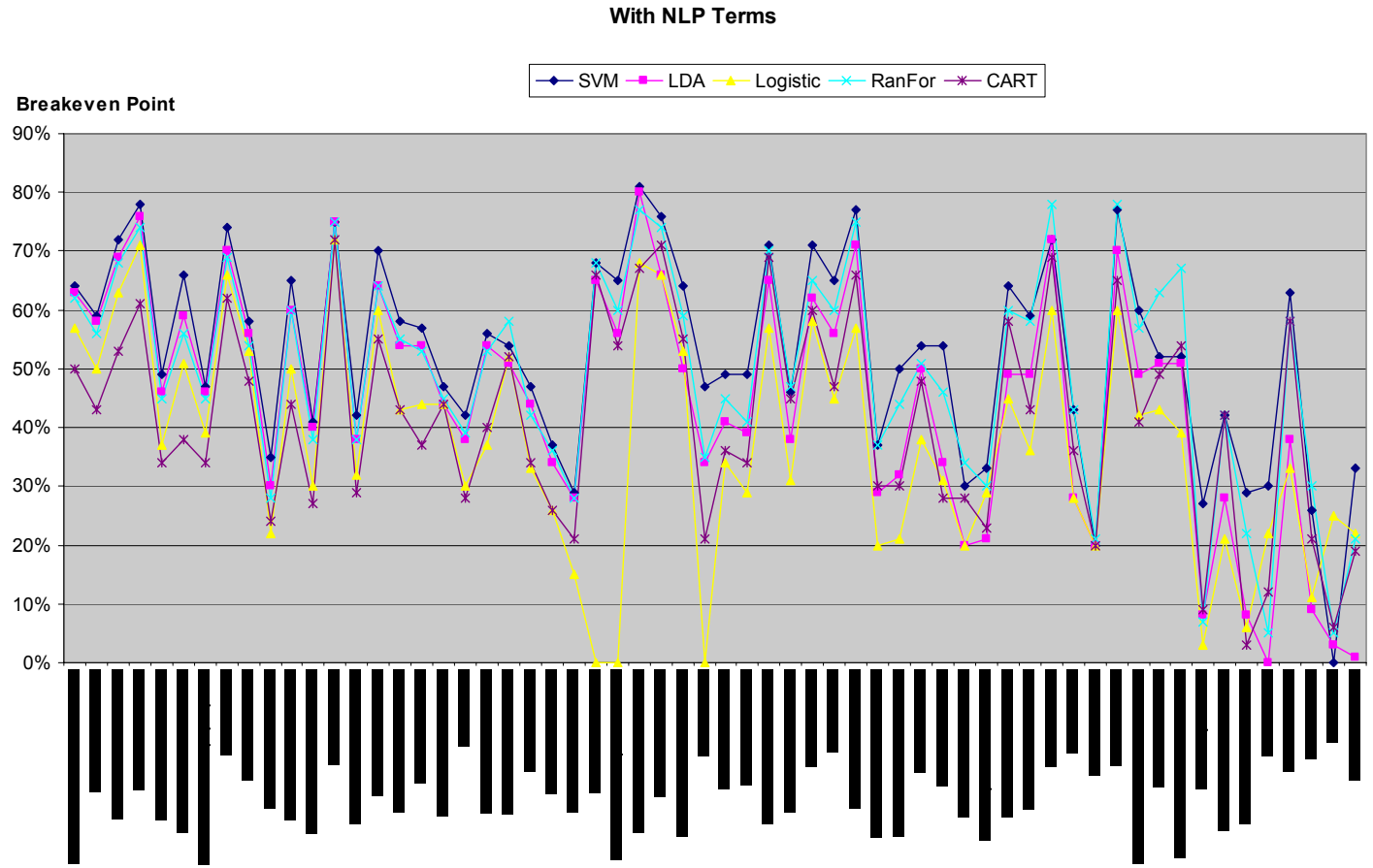


# Difference

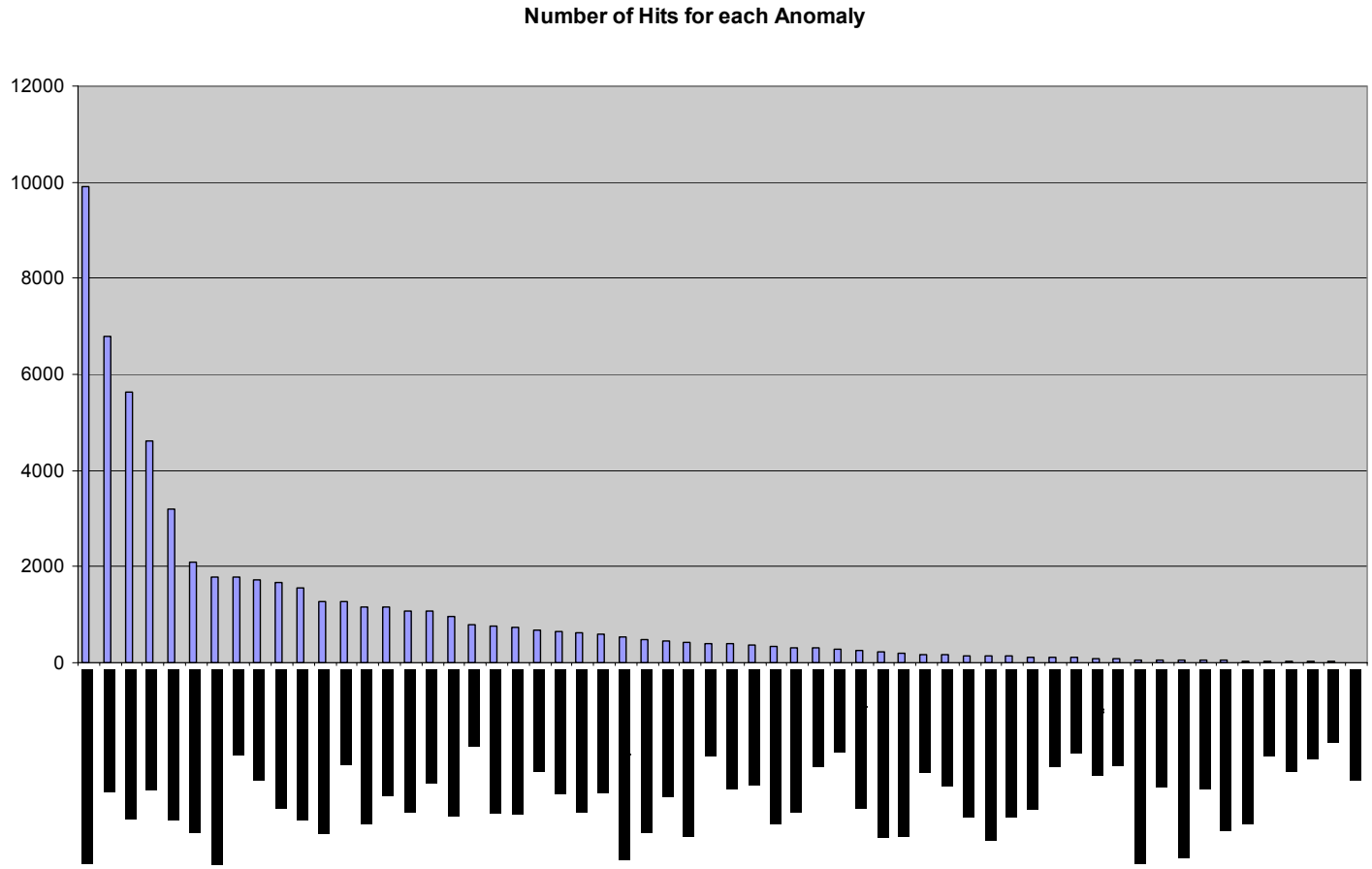
Spike is due to the fact
That SVM was not run
On these categories.

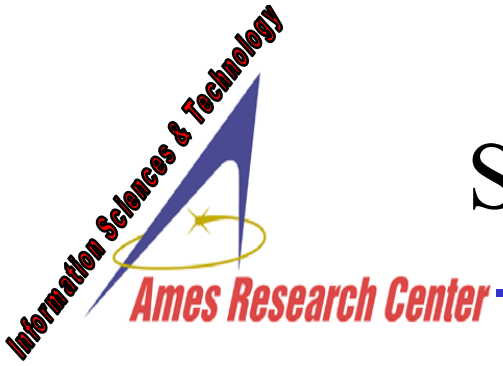


# Results with NLP

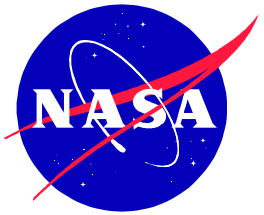


# Anomaly Frequencies






# Summary and Conclusions




- For clustering documents we find that the distributional approach suggested by Banerjee et. al. works well.
- We have discussed some reasons why vMF clustering may be useful in this application.
- We have explored the use of NLP and language normalization in detail for classification purposes.
- Results indicate no significant benefit in this classification task although the NLP methods used were extremely expensive.
- Data is available at http://ti.arc.nasa.gov/people/ashok